A statistical model for epigenetic control of miRNAs

Lisa Barros de Andrade e Sousa[1] **and Annalisa Marsico**[1,2]

[1]**Max Planck Institute for Molecular Genetics, Berlin**
[2]**Freie Universität Berlin**

ABSTRACT

microRNAs are small, non-coding RNAs involved in post-transcriptional gene regulation. Since the dysregulation of only a few miRNAs can affect many biological pathways, miRNAs are thought to play a key role in cancer development and can be used as biomarkers for cancer diagnosis and prognosis. In order to understand how miRNA dysregulation leads to a cancer phenotype it is important to determine the basic regulatory mechanisms that drive miRNA expression. Although much is known about miRNA-mediated post-transcriptional regulation, little is known about the epigenetic control of miRNAs. Here, we performed cell-line specific miRNA promoter predictions and built a classification model for expressed and non-expressed miRNAs based on several epigenetic features, e.g. histone marks and DNA methylation at both, miRNA promoters and around miRNA hairpins. We were able to classify intragenic and intergenic miRNAs with an accuracy of 79% and 85%, respectively, and identified the most important features for classification via feature selection. Surprisingly, we found that DNA methylation seems to have a dual role in regulating miRNA expression at transcriptional level: at promoters, high levels of DNA methylation correlate with transcriptional repression, while around miRNA hairpins high levels of DNA methylation have a positive impact on the expression level of the mature miRNA.

Keywords: miRNA, epigenetic regulation, regularized logistic regression, DNA methylation

INTRODUCTION

MiRNAs are small non-coding RNAs involved in many fundamental biological processes such as development, differentiation or apoptosis by means of post-transcriptional regulation of target genes, mainly via silencing mechanisms (Carleton et al., 2007; Plasterk, 2006). It is estimated that more than 50% of all human protein coding genes are regulated by miRNAs in order to increase the precision of gene expression by damping gene expression outside of optimal physiological boundaries (Herranz and Cohen, 2010). The strong conservation of several miRNAs, even among distantly related species, reflects their important role in biological processes (Iorio et al., 2010). The number of miRNAs in the human genome is about 2000 (Kozomara and Griffiths-Jones, 2014), and approximately 50% of them are located in introns or exons of other transcripts, referred to as host genes. About two third of the intragenic miRNAs are co-regulated with their host gene, while the remaining miRNAs are transcribed from independent intragenic promoters in a tissue-specific manner (Ozsolak et al., 2008). Other miRNAs are found in intergenic regions with independent transcription units (Kim et al., 2009). With few exceptions, miRNA transcription is initiated by RNA Polymerase II (Pol II) (Lee et al., 2004). The mature miRNA is derived from a large primary transcript (pri-miRNA) that is rapidly processed into a small hairpin structured precursor miRNA (pre-miRNA) by the Microprocessor. Pri-miRNAs have a short half-life because the Microprocessor cleavage mostly occurs co-transcriptionally. Hence, the primary transcript is difficult to detect with conventional sequencing techniques. The released pre-miRNA is exported to the cytoplasm by Exportin5 (Cullen, 2004). In a last step, the pre-miRNA is cleaved by Dicer, releasing a *miRNA-5p:miRNA-3p* duplex of $\sim$ 22bp length (Hutvágner et al., 2001). The duplex is loaded into the Ago2 protein to form the RNA-induced silencing complex (RISC) (Chendrimada et al., 2005). Only one strand of the duplex remains stably associated with Ago2 (the guide strand) while the other strand is degraded (Chiang et al., 2010). Figure 1 gives a general overview on the miRNA biogenesis pathway.

Since most miRNAs are transcribed by Pol II like protein coding genes, it is assumed that miRNA genes undergo similar regulatory mechanisms, including epigenetic regulation. The identification of epigenetic modifications at gene loci regions, which dynamically regulate gene expression, is an active research area. For instance, tissue-specific gene expression can be mediated through promoter silencing by methylation of histone tails or by DNA methylation of CpG sites within the promoter

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2423v1 | CC BY 4.0 Open Access | rec: 6 Sep 2016, publ: 6 Sep 2016

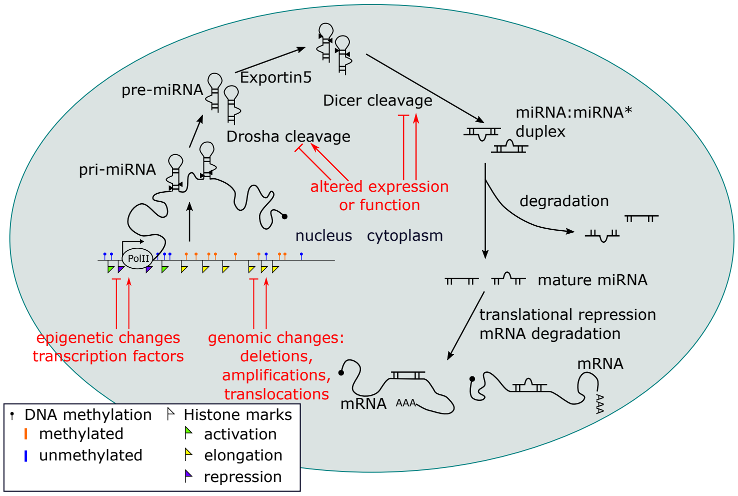

Figure 1. An overview on the miRNA biogenesis and dysregulation of miRNAs in cancer.

region. Since many human protein coding genes are also regulated by miRNAs, the dysregulation of miRNA expression is widely recognized as a hallmark of human cancer. Depending on the cellular context, miRNAs can function as tumor-suppressors or oncogenes (Melo and Esteller, 2011), and it was shown that the profile of miRNA expression is sufficient to classify a variety of human tumors or to discriminate between metastatic and non-metastatic tumors (Lu et al., 2005). Epigenetic dysregulation of miRNAs in cancer is driven by aberrant activity of chromatin remodellers, causing changes in the chromatin structure. Such structural changes can affect miRNA promoters, leading to aberrant gene activation or silencing (Jones and Baylin, 2007), or miRNA processing, therefore affecting the chromatin structure in genomic regions around the mature microRNA. Despite this general knowledge, the detailed mechanisms of aberrant miRNA regulation in cancer are still poorly understood. To facilitate the application of miRNAs as diagnostic and prognostic biomarkers in modern medicine it is important to understand the basic regulatory mechanisms that drive miRNA transcription in general, and then use this knowledge to interpret cancer data.

Karlic et al. [2010] showed the utility of machine learning techniques to uncover epigenetic regulatory mechanisms for protein-coding genes (Karlić et al., 2010). By using a linear regression model they were able to predict gene expression from the level of various histone modifications at gene promoters . Furthermore, they demonstrated that only a small number of histone modifications was necessary to accurately predict gene expression.

Following the rationale that miRNAs, like protein coding genes, are regulated at epigenetic level, we applied a regularized logistic regression model to uncover epigenetic mechanisms that drive miRNA expression. In particular, histone modification levels and DNA methylation patterns at both, miRNA promoter and regions around the miRNA precursors were used as features. A crucial step for the model construction was the identification of miRNA promoters, which is a difficult task given the poor annotation of pri-miRNA transcripts. For large-scale identification of miRNA promoters we used PROmiRNA (Marsico et al., 2013), a miRNA promoter prediction tool which supports the prediction of tissue-specific promoters, while making only few structural but no epigenetic assumptions about the similarity of protein-coding and miRNA promoters. Our model uncovered important epigenetic marks at promoter and pre-miRNA level and revealed that a subset of epigenetic features is sufficient to classify miRNAs into expressed and non-expressed with high accuracy (79% for intragenic and 85% for intergenic miRNAs).

Those results were obtained by applying the model to the Hela-S3 cell-line and could be confirmed in another cell line, Imr90, suggesting that, although epigenetic patterns are cell-line specific, the principles governing epigenetic regulation of miRNAs are conserved. Interestingly, in both cell lines we observed an inverted DNA methylation pattern for miRNA promoters and pre-miRNA flanking regions: DNA methylation correlates negatively with miRNA expression at promoter level, but positively at pre-miRNA level, suggesting the intriguing hypothesis that demethylation of promoters is necessary to activate miRNA transcription, but methylation in the regions flanking the pre-miRNA is necessary for elongation of transcription and production of the mature miRNA.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2423v1 | CC BY 4.0 Open Access | rec: 6 Sep 2016, publ: 6 Sep 2016

MATERIAL AND METHODS

Data Sets:

Human miRNAs

The annotation of human mature and pre-miRNAs was obtained from miRBase release 20 (Kozomara and Griffiths-Jones, 2014). MiRNA classification into intergenic and intragenic was downloaded from the miRIAD database version 2014 (Hinske et al., 2014).

miRNA expression

Small RNA-Seq data for Hela-S3 and Imr90 cell line were downloaded from ENCODE/Cold Spring Harbor Lab track, including two replicates. Read counts for each mature miRNA were summed over the $5'$ and $3'$ arm, averaged over both replicates and assigned to the respective pre-miRNA. The read count distribution followed approximately a bimodal distribution, which could be described by a mixture model. Given the non-Gaussian nature of the distributions we used the function npEM from the R package ”mixtools”, which accounts for non-Gaussian distributions (Derek et al., 2015), to retrieve the shape of both distributions. According to the mixture distribution, two thresholds were inferred to classify miRNAs in expressed, lowly expressed and non-expressed miRNAs. The class of lowly expressed pre-miRNAs corresponded to the read count interval where the distributions overlap and an unambiguous classification into expressed or non-expressed was not possible. Such miRNAs were excluded form further analysis.

Histone modification ChIP-Seq data

ChIP-Seq data for eleven histone modifications (see Table 1) were downloaded from ENCODE/Broad Institute track for Hela-S3 cell line and from the Epigenome Roadmap [July 2011] for Imr90 cell line. Reads of each histone modification were mapped to the 500bp region surrounding the pre-miRNA and to the 1000bp region surrounding the corresponding predicted promoters. We added an additional pseudocount of one (to avoid undefined values of the logarithm when read count equals zero) to the read count and normalized by log-transformation followed by z-score transformation.

Histone Modification	Putative Function	Location
H2AZ	transcriptional activation	promoter
H3K27ac	transcriptional activation	promoter, enhancer
H3K27me3	transcriptional repression	promoter
H3K36me3	transcription elongation	gene body
H3K4me1	transcriptional activation	enhancer
H3K4me2	transcriptional activation	promoter, enhancer
H3K4me3	transcriptional activation	promoter
H3K79me2	transcription elongation	gene body
H3K9ac	transcriptional activation	promoter
H3K9me3	transcriptional repression	promoter, enhancer, gene body
H4K20me1	transcriptional activation	gene body

Table 1. The histone modifications used for the logistic regression model and their gene regulatory function (Zhou et al., 2011).

DNA methylation Bisulphite-Seq data

The coordinates and β-values of CpG sites from Hela-S3 and Imr90 cell line were downloaded from ENCODE Methyl 450K Bead Array track. β-values range from 0 to 1 where 0 to 0.2, 0.2 to 0.6 and 0.6 to 1 represent an unmethylated, partially methylated and methylated CpG site, respectively. The median of β-values from CpG sites that mapped to the 500bp region surrounding the pre-miRNA determined the methylation status of the pre-miRNA region. The methylation status of the predicted miRNA promoter regions was calculated within a consensus region surrounding the promoter. Therefore we started with an initial interval of 100bp surrounding the centre of the promoter and extended this interval upstream and downstream by a 100 bp sliding window if the median β-value of CpG sites within the next sliding window did not deviate more than 0.3 from the median value within the current region. The maximum size of this consensus region was 2000bp. The methylation status of promoter regions was then set to the mean β-value from the CpG sites inside the consensus region. We normalized the methylation status by applying z-score transformation.

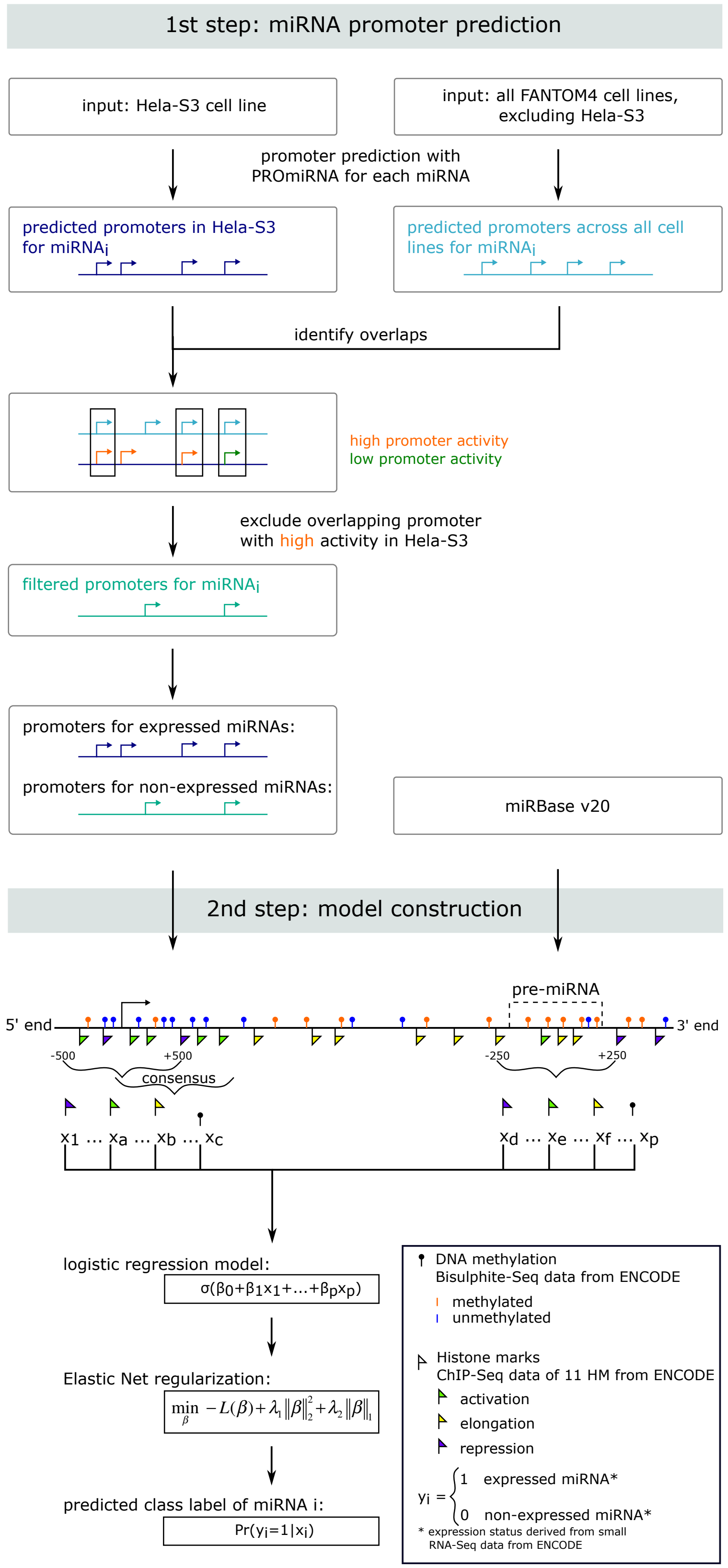
1st step: miRNA promoter prediction
input: Hela-S3 cell line
input: all FANTOM4 cell lines, excluding Hela-S3
promoter prediction with PROmiRNA for each miRNA
predicted promoters in Hela-S3 for miRNAi
predicted promoters across all cell lines for miRNAi
identify overlaps
high promoter activity
low promoter activity
exclude overlapping promoter with high activity in Hela-S3
filtered promoters for miRNAi
promoters for expressed miRNAs:
promoters for non-expressed miRNAs:
miRBase v20
2nd step: model construction
pre-miRNA
5' end
3' end
-500
+500
consensus
-250
+250
x1 … xa … xb … xc
xd … xe … xf … xp
logistic regression model:
σ(β0+β1x1+…+βpxp)
Elastic Net regularization:
min_β −L(β)+λ1‖β‖²₂+λ2‖β‖₁
predicted class label of miRNA i:
Pr(yi=1|xi)
DNA methylation
Bisulphite-Seq data from ENCODE
methylated
unmethylated
Histone marks
ChIP-Seq data of 11 HM from ENCODE
activation
elongation
repression
yi = 1 expressed miRNA*
0 non-expressed miRNA*
* expression status derived from small RNA-Seq data from ENCODE

Figure 2. Workflow to build the miRNA classification model.

Workflow:

We set up a workflow to model the impact of epigenetic marks on miRNA expression (Figure 2). The workflow is composed of two main parts: 1) miRNA promoter prediction and 2) model construction.

The first step, the miRNA promoter prediction, is crucial to identify the region where promoter features are computed. We used the PROmiRNA tool to predict promoters of active miRNAs in both, Hela-S3 and Imr90 cell lines. PROmiRNA uses deepCAGE read coverage together with other sequence feature to predict miRNA promoters versus background. It provides predictions for all three types of miRNA promoters – host gene, intergenic and intronic promoters – and is sensitive enough to capture transcription start sites of lowly expressed miRNA genes due to a prior probability distribution which does not depend on the miRNA expression. In addition, PROmiRNA, unlike other tools, does not use epigenetic features for promoter predictions, allowing an unbiased use of such features in the classification model. For the classification of miRNAs in expressed and non-expressed, it is necessary to predict the promoters of expressed as well as non-expressed miRNAs. PROmiRNA predicts a set of active promoters in cell line X for expressed miRNAs. However, promoters of non-expressed miRNAs are inactive and hence, no CAGE tags – the basis for promoter prediction in PROmiRNA – can be observed for those promoters. To overcome this problem, we exploited a special feature of PROmiRNA that allowed us to use multiple cell lines as input to predict alternative miRNA promoters across those cell lines. Since most miRNAs are differentially expressed across different cell lines, and miRNAs being non-expressed in cell line X are potentially expressed in other cell lines, the prediction across multiple cell lines can be used to impute inactive promoters for non-expressed miRNAs. To ensure that all the promoters are inactive we excluded promoters overlapping with highly active promoters in cell line X.

A set of active promoters in the Hela-S3 cell line was obtained by training PROmiRNA on Hela-S3 CAGE data from the FANTOM5 Consortium. The set of alternative promoters was predicted on all FANTOM4 libraries except the Hela-S3 library. The same procedure was employed in the case of Imr90 cells. At the end, all predicted promoters within a distance of at most 10bp to each other were merged to one promoter region and the central 1000bp region was considered as the new promoter region.

In the second step, a regularized logistic regression model based on epigenetic features was built to predict expressed vs non-expressed miRNAs. To control fluctuations in expression values introduced by miRNA biogenesis processing steps we decided to discretize the response variable of the model, i.e. the expression level of each mature miRNA, and used a logistic regression instead of a linear regression. A classification model revealed to be more robust and accurate than a regression model for this task.

The feature set for the logistic regression model comprises the normalized levels of eleven histone modifications (see Table 1) and methylation status at promoter and pre-miRNA regions. To incorporate miRNA and promoter regions without methylation status (due to missing β-values for CpG sites) into the model, we imputed the missing methylation status by randomly sampling a β-value between 0.2 and 0.6, which represents a partial methylation.

The logistic regression model was regularized with Elastic Net, to account for correlated features and to perform feature selection. The regularized logistic regression model was fitted in R with the *glmnet* function from the R package *glmnet*. An Elastic Net Mixing parameter of $\alpha = 0,5$ was arbitrarily chosen to build our model. The model was fitted on a training set, containing a subset of the expressed miRNAs (full model: 1000 of 1833; intragenic model: 700 of 895; intergenic model: 500 of 938 data points, keeping in mind that there can be multiple promoter predictions per miRNA) and non-expressed miRNAs (full model: 1000 of 1965; intragenic model: 700 of 1287; intergenic model: 500 of 678 data points), drawn at random without replacement. The regularization path is computed at a grid of lambda values. The optimal lambda value is obtained by choosing the lambda value that maximizes the model auc. Here, we did not use the cross-validation function provided by the *glmnet* package because we wanted to obtain the lambda value that maximizes the balanced class accuracy, enabling a good prediction of expressed and non-expressed miRNAs. To optimize the final model, we chose a final value for lambda from a vector of optimal lambda values from 200 bootstrap runs (see algorithm 1).

All performances measures were computed on an independent test set consisting of the full set without the training set, using $elnet_{final}$. The accuracy of the regularized logistic regression model was balanced by taking the mean of the true positive rate and false negative rate. In order to take into account the stability of the model, the final model accuracy and shown feature coefficients are averaged over 500 runs.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2423v1 | CC BY 4.0 Open Access | rec: 6 Sep 2016, publ: 6 Sep 2016

Algorithm 1 Training and testing of the regularized logistic regression model.

```
function GET_ACC($elnet, data\_set$)
    for $j$ in 1 : $length(\lambda\_seq)$ do
        $TPR$ = fraction of correctly predicted expressed miRNAs on $data\_set$ at $\lambda_j$
        $TNR$ = fraction of correctly predicted non-expressed miRNAs on $data\_set$ at $\lambda_j$
        $acc[j] = mean(TPR, TNR)$
    end for
    return acc
end function
$training\_set$ = random sampling without replacement from $full\_set$
$test\_set = full\_set - training\_set$
for $i$ in 1 : 200 do
    $training\_set_i$ = random sampling without replacement from $training\_set$
    $validation\_set_i = training\_set - training\_set_i$
    $elnet_i$ = Elastic Net model fitted on $training\_set_i$
    $acc_i$ = GET_ACC($elnet_i, validation\_set_i$)
    $\lambda_{opt}[i] = \lambda$ value corresponding to the maximum $acc_i$
end for
$training\_set_{final}$ = random sampling without replacement from $training\_set$
$validation\_set_{final} = training\_set - training\_set_{final}$
$elnet_{final}$ = Elastic Net model fit on $training\_set_{final}$ and $\lambda_{opt}$
$acc_{final}$ = GET_ACC($elnet_{final}, validation\_set_{final}$)
$\lambda_{final} = \lambda$ value corresponding to the maximum $acc_{final}$
$selected\_features$ = features of $elnet_{final}$ at $\lambda_{final}$
$accuracy$ = fraction of correctly predicted miRNAs on $test\_set$ with $elnet_{final}$ at $\lambda_{final}$
```

RESULTS

In this section we present the results from the Elastic Net model trained on histone modifications and DNA methylation in two different cell lines (Hela-S3 and Imr90) and discuss the most important features that seem to drive miRNA expression.

Model for miRNA expression in Hela-S3 cells

A correlation analysis for promoter and pre-miRNA features showed a strong correlation between different features. In particular, the histone modifications that are putatively responsible for transcriptional activation (H2AZ, H3K27ac, H3K4me2, H3K4me3 and H3K9ac) correlate positively with each other and correlate negatively with DNA methylation at promoter and pre-miRNA level. Parameter estimates with a logistic regression model can be unstable in the presence of collinearity. For this reason, we chose to use the Elastic Net regularization because it combines the Lasso and Ridge shrinkage penalties to perform feature selection in combination with a grouped selection of correlated features by which groups of correlated features are assigned similar, and numerically more stable, coefficient values (Zou and Hastie, 2005). Therefore, Elastic Net is able to generate a sparse model while handling feature collinearity. Compared to other models that can handle feature collinearity, Elastic Net provides a straight forward interpretation of the features.

In order to analyse the impact of different epigenetic marks at pre-miRNA and promoter level, we build a full model on all pre-miRNA and promoter features (twenty-two features for the eleven histone marks of pre-miRNA and promoter region; two features for DNA methylation of pre-miRNA and promoter region). The model reached a total accuracy of 78% on an independent test set, while predicting expressed miRNAs as good as non-expressed miRNAs. Regularizing the logistic regression model helped to identify important features that have an impact on the classification of miRNAs into expressed and non-expressed. Feature selection revealed that DNA methylation of pre-miRNA and promoter region, elongation marks H3K79me2 and H3K36me3 at pre-miRNA and promoter level, and repressive mark H3K27me3 at promoters are important features for the classification (see Figure 3 b)). DNA methylation seems to play a crucial role at pre-miRNA, as well as promoter level, but since the sign of the feature coefficients is inverted, the function of DNA methylation seems to differ between both regions. The DNA methylation feature has a negative coefficient for the promoter region, suggesting a repressive function for DNA methylation at miRNA promoters. In contrast, high levels of DNA methylation around miRNA precursors correlate positively with expression, indicating that DNA methylation plays a different role at pre-miRNA than at promoter level (see Figure 4). The repressive mark H3K27me3 seems to have a higher impact on the classification than any of the active marks that are found at promoter level. It has a negative feature coefficient whereas the elongation marks have positive feature coefficients, as expected for actively transcribed regions. Histone 3 lysine 4 methylations are a general mark of active transcription at promoter level. However, the coefficient of H3K4me3 at promoter level is negative. As this was not expected, we attributed this to putative differences in the epigenetic landscape between intergenic and intragenic miRNAs, as discussed further in the Discussion section of the paper.

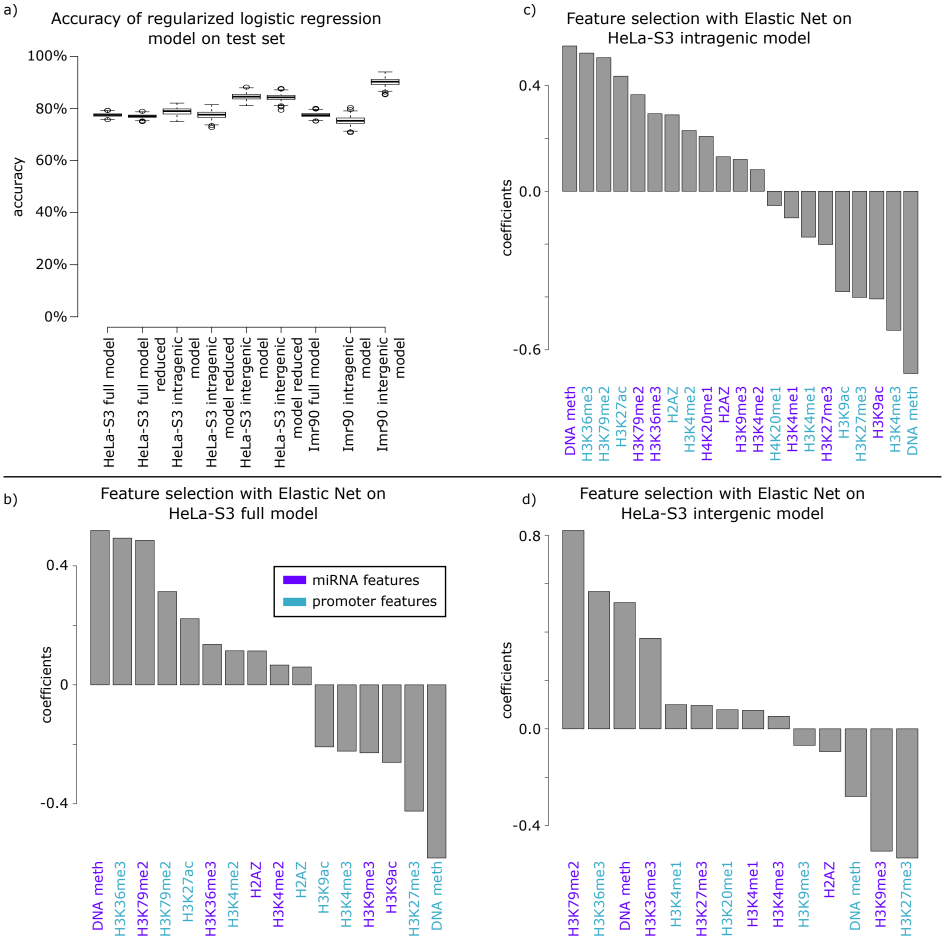

Figure 3. Model Performance and feature selection, not showing features that were shrunk to zero. a) accuracy of the different regularized logistic regression models measured on the test set. Feature selection was performed with Elastic Net on Hela-S3 b) full model c) intragenic model and d) intergenic model.

As epigenetic patterns of host genes could influence the classification results for intragenic miRNAs, we built two separate sub-models, one based on intragenic and the other one based on intergenic miRNAs. The division of the full model into the two sub-models improved the model performance, increasing the accuracy to 79% and 85% for the intragenic und intergenic miRNAs, respectively. Figure 3 c) - d) shows that the feature selection with Elastic Net was more effective for the intergenic than for the intragenic miRNA model, i.e. more features are shrunk to zero (3 for intragenic compared to 10 for intergenic). This indicates that the host gene epigenetic marks introduce noise into the intragenic model, impeding a clear identification of the important features for the miRNA itself. This might also explain the higher accuracy of the intergenic model. In addition, different feature sets were selected by the two models suggesting that the epigenetic landscape of intragenic and intergenic miRNAs might differ. The strongest features for the intergenic miRNAs are elongation and repressive marks. DNA methylation at promoter and pre-miRNA level plays a less important role in the intergenic versus intragenic miRNAs, but the inverted methylation pattern from promoter to pre-miRNA is still observed for both miRNA classes, indicating that this observation is not a bias introduced by host gene epigenetic marks. A clear difference can be seen in the role of H3K4me3, which is among the important negative features in the intragenic model, while its coefficient in the intergenic model is shrunk to zero.

Progressive feature addition

To evaluate if a subset of important features is sufficient to classify miRNAs into expressed and non-expressed miRNAs, we performed a progressive feature addition for the full, intragenic and intergenic model. Therefore, we progressively added features (trained on a training set) that most improved the accuracy (evaluated on a test set) of the model until we reached +/- 1% of the accuracy obtained by the model with all features. The progressive feature addition showed that only a subset of features (6, 10 and 7 features for the full, intragenic and intergenic model, respectively) is sufficient to reach a comparable model performance (see Figure 3 a)). Interestingly, H3K4me3 is not any longer among the selected features for the full and intragenic model. The elongation and repressive marks as well as DNA methylation are among the selected features in all three models. Hence, these three types of features – elongation, repression and DNA methylation – are most informative for miRNA classification.

Model verification on Imr90 cells

To verify whether the results hold for other cell lines, we built the same regularized logistic regression model with the same epigenetic features for the Imr90 cell line. The full model reaches the same accuracy compared to the model build on Hela-S3 (Figure 3). The performance of the intergenic model could be increased by 5%, while the performance of the intragenic model decreased by 4%. In addition, the elongation marks are among the important features, and the same DNA methylation pattern was observed at miRNA transcripts from the promoter to the pre-miRNA region. However, the repressive marks are not as prominent as in the Hela-S3 models. Overall this analysis showed comparable performance and similar important features in two different cell lines.

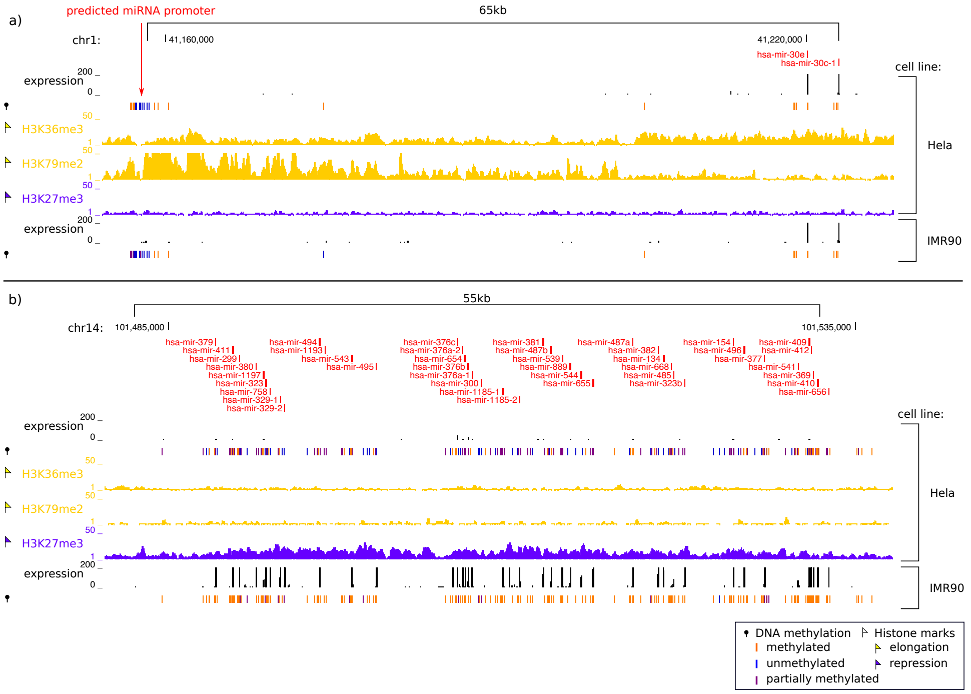

Figure 4. Genome Browser examples. a) shows an unmethylated promoter region and a methylated miRNA gene body for two expressed miRNAs in both cell lines, Hela-S3 and Imr90. b) shows a cluster of miRNAs which is expressed in Imr90 cell line and not expressed in Hela-S3 cell line. It can be observed that the miRNA body region is heavily methylated in the Imr90 cell line while much less methylation is observed in the Hela-S3 cell line.

DISCUSSION

The goal of this study was to investigate the role of epigenetic marks in miRNA expression, using a logistic regression model. A major challenge was the prediction of cell line specific promoters not only for expressed but also for non-expressed miRNA, which have inactive promoters that cannot be detected by simply looking at tissue-specific CAGE read enrichment. To take care of this we adapted the PROmiRNA software to predict promoters in cell lines other than the cell line of interest and used these predictions to infer the location of putative promoters. Although we cannot guarantee that the promoter location in one cell line where the miRNA is active is the same in the cell line where the the miRNA is inactive, this method showed to work well in practice.

We discretized miRNA expression in our model, and preferred a logistic regression over a linear regression, to make it more robust to several sources of noise which affect miRNA expresison levels. Although even discretizing the response does not fully account for biases in the expression levels due to steps of the miRNA biogenesis pathway not explicitly modelled in our classifier, a logistic regression model was more robust than a linear regression model, returning more stable coefficients and higher accuracy. Our results indicate that miRNAs, like protein-coding genes, are regulated at epigenetic level and their expression class can be predicted with high accuracy and from a handful of features. In detail, the accuracy was 79% for intragenic and 85% for intergenic miRNAs. Expectedly, elongation marks, associated to active transcription, are important features to distinguish expressed from non-expressed miRNAs for both intragenic and intergenic miRNAs, while the repressive mark marks promoters of inactive miRNAs.

DNA methylation status at pre-miRNA and promoter level, elongation marks at pre-miRNA and promoter level, and repressive marks at promoter level revealed to be the most important features for classification. As expected, DNA methylation at promoters negatively regulates promoter activation, as already observed for protein-coding genes. Interestingly, the positive DNA methylation coefficient at pre-miRNA level suggests that DNA methylation in the miRNA primary transcript gene body promotes miRNA transcription. Examples of miRNAs exibiting this pattern are depicted in Figure 4. While the function of DNA methylation seems to differ between promoters and pre-miRNA regions, the biological reason for that is not yet clear. A study from Baubec et al. showed that gene body methylation is associated to H3K36me3 in actively transcribed genes, suggesting that gene body methylation might be responsible for proper splicing (Baubec et al., 2015). In miRNAs, gene body methylation might be responsible for alternative promoter silencing of intragenic miRNAs, inhibiting spurious transcription initiation, which could explain the higher importance of this feature for the intragenic model.

H3K4me3, which is a general mark of active transcription at promoter level, has a negative coefficient in the intragenic model. Intragenic miRNAs can be transcribed from the host gene promoter but also exhibit alternative promoters. To build the logistic regression model we used host gene, as well as alternative predicted miRNA promoters, with no possibility to distinguish which one is driving miRNA transcription in the specific cell line. In a case where the host gene is expressed but the miRNA is non-expressed, the epigenetic marks of the host gene introduce a bias into the model. This explains why H3K4me3 has a negative coefficient in the intragenic model but not in the intergenic model where no host gene epigenetic marks can interfere with miRNA expression. In the intergenic miRNA model the H3K4me3 feature is irrelevant for classification (i.e coefficient shrunk to zero). Bivalent promoters are marked by the repressive H3K27me3 and activating H3K4me3 mark simultaneously, which leads to gene repression (Bernstein et al., 2006). Hence, the active mark H3K4me3 is also found at promoters of non-expressed miRNAs, explaining why it cannnot be used to distinguish expressed from non-expressed miRNAs.

Applying the model on the Imr90 cell line leads to similar classification rates and feature selection results. This highlights the fact that, although miRNA expression is cell-line specific, the epigenetic rules that govern miRNA expression are general and conserved among cell lines.

CONCLUSION

This study shows the importance of statistical models to reveal underlying patterns of epigenetic gene regulation. Our analysis discovered an unexpected role for DNA methylation in pre-miRNA regions to promote or mantain miRNA transcription. Further analysis have to show whether DNA methylation is just a by-product of transcription possibly inhibiting spurious miRNA transcription initiation or whether specific CpG sites have additional functions such as controlling alternative promoters or mediating the binding of specific factors.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2423v1 | CC BY 4.0 Open Access | rec: 6 Sep 2016, publ: 6 Sep 2016

REFERENCES

Baubec, T., Colombo, D. F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A. R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546):243–7.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–26.

Carleton, M., Cleary, M. A., and Linsley, P. S. (2007). MicroRNAs and cell cycle regulation. *Cell cycle (Georgetown, Tex.)*, 6(17):2127–32.

Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436(7051):740–4.

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., Johnston, W. K., Russ, C., Luo, S., Babiarz, J. E., Blelloch, R., Schroth, G. P., Nusbaum, C., and Bartel, D. P. (2010). Mammalian microRNAs: Experimental evaluation of novel and previously annotated genes. *Genes and Development*, 24:992–1009.

Cullen, B. R. (2004). Transcription and processing of human microRNA precursors. *Molecular cell*, 16(6):861–5.

Derek, A., Hunter, D., Elmore, R., Hettmansperger, T., and Thomas, H. (2015). Package ‘ mixtools ’.

Herranz, H. and Cohen, S. M. (2010). MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. *Genes & development*, 24(13):1339–44.

Hinske, L. C., França, G. S., Torres, H. A. M., Ohara, D. T., Lopes-Ramos, C. M., Heyn, J., Reis, L. F. L., Ohno-Machado, L., Kreth, S., and Galante, P. A. F. (2014). miRIAD-integrating microRNA inter- and intragenic data. *Database : the journal of biological databases and curation*, 2014(0):bau099–.

Hutvágner, G., McLachlan, J., Pasquinelli, A. E., Bálint, E., Tuschl, T., and Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science (New York, N.Y.)*, 293(5531):834–8.

Iorio, M. V., Piovan, C., and Croce, C. M. (2010). Interplay between microRNAs and the epigenetic machinery: An intricate network. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1799(10-12):694–701.

Jones, P. A. and Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, 128(4):683–92.

Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107:2926–2931.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology*, 10(2):126–39.

Kozomara, A. and Griffiths-Jones, S. (2014). mirbase: annotating high confidence micrornas using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20):4051–60.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., and Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–8.

Marsico, A., Huska, M. R., Lasserre, J., Hu, H., Vucicevic, D., Musahl, A., Orom, U. A., and Vingron, M. (2013). PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome biology*, 14(8):R84.

Melo, S. a. and Esteller, M. (2011). Dysregulation of microRNAs in cancer: Playing with fire. *FEBS Letters*, 585(13):2087–2099.

Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., Zhang, X., Song, J. S., and Fisher, D. E. (2008). Chromatin structure analyses identify miRNA promoters. *Genes and Development*, 22:3172–3183.

Plasterk, R. H. A. (2006). Micro RNAs in animal development. *Cell*, 124(5):877–81.

Zhou, V. W., Goren, A., and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics*, 12(1):7–18.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*.